한국컴퓨터정보학회 하계학술대회 논문집 제19권 제2호 (2011. 6)

특허정보검색 알고리즘에 관한 연구

고광수°, 정원교*, 신영근*, 박상성*, 장동식*
°*고려대학교 산업경영공학과
e-mail: take0815@korea.ac.kr

A Study on The Patent Information Retrieval Algorithm

Gwang Su Go°, Won Gyo Jung*, Young Geun Shin*, Sang Sung Park*, Dong Sik Jang*
°*School of Industrial Management Engineering, Korea University

● 요 약 ●

본 연구에서는 특허문서에 사용된 핵심키워드 찾아내고 추출된 핵심키워드에 가중치를 부여하여 특허데이터DB에서 질의문서와 유사한 특허기술문서를 찾고 유사도 순으로 우선 배치하여 검색에 효율을 높일 수 있는 알고리즘을 제안한다. 본 연구는 제안한 알고리즘은 검색결과에 대하여 질의한 문서와 유사한 문서 순으로 랭크가 가능하기 때문에 검색의 효율을 높이는 효과를 가지고 온다.

키워드: Precision, Patent, Retrieval

I. 서론

현재 선행기술조사는 찾고자하는 특허문서에 대한 핵심키워드를 먼저 알아내고 그 키워드를 이용한 데이터 검색작업이 이루어지고 있다. 일반적으로, 데이터 검색작업은 확장검색키워드 추출, 확장검색식 작성, 결과데이터 검토, 추가 확장검색키워드 추출과 같은 단계를 반복적으로 수행한다. 이러한 수십에서 수백건에 이르는 데이터를 검토하는 작업은 많은 노력과 시간이 소요되게 된다. 특허 검색 사용자가 원하는 특허를 찾기 위해서는 효과적인 검색식을 만드는 것이 중요하다.[1] 하지만 사용자가 효과적인 검색식을 만들어내기 위해서는 해당 기술에 대한 핵심키워드를 찾아낼 수 있어야 함은 물론 찾아낸 키워드를 적절히 사용하여 검색식을 만들어 내는 능력이 필요하다.

본 논문은 특허문서의 검색결과가 사용자가 찾고자하는 문서와 유사한 순으로 검색되어지는 알고리즘을 연구하는데 목적이 있다. 즉, 본 연구에서는 검색된 결과 데이터들에 대해서 질의문서와의 유사정도의 정보를 추가적으로 제공하여 관련 기술을 분류하는 결과검토 작업에 소요되는 시간과 노력을 줄이고자 한다.

본 연구에서는 특허문서의 색인어들에 대한 가중치를 부여하는 TF-IDF(Term Frequency Inverse document Frequency)알고리즘을 이용하여, 질의문서와 유사한 순으로 검색결과를 나타낸다. 실험데이터로는 반도체 웨이퍼(Wafer), 터치(Touch), LTE(4세대 통신)기술에 대한 데이터를 이용한다.

검색결과에 대한 정확도(Precision) 실험을 통해 알고리즘 성능을 측정한다. 본 연구는 서론에 이어 제2장에서는 관련 연구를 기술하였고, 제3장에서는 제안된 알고리즘 설명을 하였으며 그리고 제4장에서는 실험 및 결과를 다루고, 마지막으로 제5장에서는 결론 및 향후과제를 기술한다.

II. 관련 연구

현재, 유사특허기술문서를 효율적으로 검색하는 연구로 유사키워드 제안에 대한 연구와 키워드 정보를 이용한 문서분류에 대한 연구가 대표적이다. 질의 검색어를 대체할 수 있는 유사키워드 후보를 제안하는 연구[2]는 “문장 내에서 함께 쓰이는 단어들이 비슷한 두 단어는 서로 비슷한 의미를 지닐 것이다”라는 가정 하에서 유사키워드 후보를 탐색한다. 유사키워드를 추출하는 단계는 먼저 문서별 중요 단어 선정을 한다. 다음으로 연관단어 뭉치를 생성하고, 생성된 연관단어 뭉치의 유사도를 계산하여 대체어 후보 목록을 생성한다. 마지막으로 대체어 수위 보정을 거쳐 유사키워드 후보를 제안한다. 다음으로 영역(Category)별 문서 분류 연구[3]는 영역별 개념 지식을 사용하여 입력 받은 문서가 어떤 영역에 속하는 지를 결정해 준다. 문서분류과정은 먼저, 영역별 개념 지식을 기구축된 문서의 집합으로부터 제목과 내용에 기반한 앵커 텍스트(anchor text)를 이용하여 개념을 보유한 키워드를 추출한다. 다음으로 추출된 키워드를 형태소 분석을 통해 색인어로 추출하고 추출된 색인어에 대해 가중치를 부여하여 영역별 개념 기반 색인어와 색인을 구축한다. 구축된 영역별 개념 기반 색인을 이용하여 새로운 웹 문서에 대해서 어떤 영역에 해당하는 가를 결정하는 자동 분류 알고리즘을 수행하게 되며 수행된 문서는 영역별로 정리된다. 이와 같은 영역별 문서 분류 연구는 웹 문서를 대상으로 연구가 이루어지고 있다.

III. 제안된 알고리즘

특허 문서 분석을 위해서는 문자로 표현된 특허문서를 분석이 가능한 형태로 변환하는 과정인 텍스트마이닝 과정이 요구된다.[4] TF-IDF는 문서에 포함된 단어들에 가중치 부여를 통하여 문서의 핵심단어를 추출해내는 용도로 사용된다. 즉, 정보 검색과 텍스트 마이닝에서 이용되는 가중치이다. 여기서 TF(단어빈도수, Term frequency)는 문서 내에 출현하는 모든 단어를 대상으로 각 단어들이 해당 문서에서 출현하는 빈도를 나타내는 값이다. TF값이 높을수록 해당문서에서 특정단어가 차지하는 중요도가 높다고 할 수 있다. DF(문서 빈도수, Document frquency)는 문서 집합에서 출현하는 단어의 빈도를 나타낸다.[5, 6] DF 값이 높을수록 해당단어는 문서 집합내에서 흔하게 사용되어진 단어라고 할 수 있다. 같은 단어일지라도 DF값은 문서집합들의 특성에 따라서 달라질 수 있다. 텍스트마이닝에서 이용되는 가중치는 하나의 특허문서 내에 존재하는 단어의 빈도값(TF)와 특허문서집합내에서 등장하는 문서 빈도수(DF)를 모두 고려하여 사용되어지므로, DF의 역수인 IDF(inverse document frequence)로 주로 이용된다. 즉, 특허문서에 출현하는 단어의 빈도값(TF)과 문서빈도수의 역수값(IDF)를 곱한 값을 이용하여 해당 특허문서의 벡터값을 표현한다. 식(1~3)은 단어의 빈도(TF)와 문서빈도수의 역수(IDF)를 계산하는 방법을 나타낸다. [7]

$$tf_{i,j} = \frac{f_{i,j}}{\sum_m f_{i,j}} \qquad \text{식(1)}$$

$f_{i,j}$ =nober of occurrences of the considered term(t_i) in document d_j

$\sum_m f_{i,j}$=the sum of number of occurrences of all terms in document d_j

$$idf_{i,j} = \log\frac{|D|}{|d : t_i \in d|} \qquad \text{식(2)}$$

|D|:문서의 총 개수

$|t \in d|$: Term을 포함한 문서의 개수

$$\text{가중치} = (tf - idf)_{i,j} = tf_{i,j} \times idf_i \qquad \text{식(3)}$$

IV. 실험 및 결과

본 논문에서는 미국에 출원되어 있는 반도체 웨이퍼(Wafer), 터치(Touch), LTE(제4세대통신) 기술에 관한 특허문서를 이용하여 실험 하였다. 검색대상은 2001년 3월에 공개되어 있는 특허문서를 대상으로 하였으며, 검색은 특허 검색사이트인 WIPS (http://www.wips.co.kr)에서 국제특허분류(IPC코드) 및 기술별 핵심키워드를 이용하여 특허기술문서 집합을 검색하였다. 특허기술 문서 집합을 만들기 위하여 수집한 특허기술문서는 반도체 웨이퍼(Wafer)에 관한 특허문서 105개, 터치(Touch)에 관한 기술문서 105개, LTE에 관한 기술문서 100개를 임의 추출 방식을 사용하여, 3개의 기술 군으로 이루어진 총 310개의 특허기술문서 집합을 생성하였다. 질의문서는 특허기술문서집합에 사용되지 않는 반도체 웨이퍼기술, 터치기술, 그리고 LTE에 관한 각각의 특허기술문서를 Query(q)로 사용하였다.

q1은 반도체 웨이퍼, q2는 터치, q3는 LTE에 관한 Query(q)를 나타낸다. 표 1.은 상위 100개에 대한 기술 분류 정확도를 나타낸다.

$$\text{정확도} = \frac{\text{유사문서갯수}}{\text{기술군별문서총갯수}} \times 100 \qquad \text{식(2)}$$

식(2)를 이용하여 각 질의에 대한 정확도를 계산하였다.

표 1. 상위100개에 대한 기술 분류 정확도

	질의문서	유사 (개수)	비유사 (개수)	정확도	정확도 평균
Screening	q1	100	0	100%	100%
	q2	100	0	100%	
	q3	100	0	100%	
TF-IDF	q1	65	35	65%	65.6%
	q2	62	38	62%	
	q3	70	30	70%	

질의문서와 동일한 기술군에 해당하는 문서의 경우는 유사한 문서로 판단하였으며, 다른 기술군에 해당하는 문서의 경우는 비유사한 문서로 판단하였다. 상위 100개의 검색결과에서 q1을 질의한 경우 반도체 웨이퍼와 관련된 문서가 65개, 터치 및 LTE에 관한 문서가 35개로 65% 정확도를 보였다. q2을 질의한 경우 터치기술과 관련된 문서가 62개, 반도체웨이퍼 및 LTE에 관한 문서가 38개로 62% 정확도를 보였다. q3을 질의한 경우 LTE와 관련된 문서가 70개, 터치 및 LTE에 관한 문서가 30개로 70% 정확도를 보였다. 스크리닝 방법의 경우 검색정확도가 100%의 검색결과를 보이지만 데이터 검색 및 추출시간이 많이 소요되는 특징을 지닌다.

표 2.는 반도체 웨이퍼 기술에 관한 데이터구성을 나타낸다.

표 2. 반도체웨이퍼 특허기술문서 집합

영역 (Category)	반도체웨이퍼 기술에 관한 특허기술문서 집합
Cleaning	35
Heating	35
Testing	30
총 합	100

반도체 웨이퍼 특허기술문서집합에 포함되어 있는 문서를 세부기술군 별로 분류하여 총 100개의 문서 중에서 반도체 웨이퍼 클리닝(Cleaning)기술에 관한 기술문서 35개, 웨이퍼 예열(Heating)에 관한 기술 35개, 웨이퍼 테스팅(Testing)에 관한 기술 30개를 사용하였다. 질의문서로 사용된 Query은 반도체 웨이퍼 클리닝(Cleaning) 기술에 관한 특허문서를 사용하였다. 표 3.는 반도체 웨이퍼 클리닝(Cleaning)기술을 Query로 사용하여 결과를 유사한 세부기술 순으로 랭킹한 정확도를 나타낸 것이다.

표 3. 반도체 웨이퍼 클리닝(Cleaning) 세부기술 분류 정확도

구분	세부기술			정확도
	Cleaning	Heating	Testing	
상위10개	7	2	1	70%
상위20개	13	4	3	65%
상위30개	19	6	5	63%

정확도 결과비교를 위하여 질의문서와의 유사도 순으로 랭킹한 결과를 상위 10개, 20개, 30개로 나누어 정확도를 비교해 보았다, 각각 70%, 65%, 63%를 정확도를 나타내었으며, 상위에 랭킹된 문서일수록 질의문서와의 유사 정확도가 높은 것으로 나타났다.

V. 결론

본 논문에서는 TF-IDF 알고리즘을 이용하여 특허기술문서집합에서 질의문서와 유사한 순으로 검색 결과를 랭킹(Ranking)하여 유사 정도에 대한 정보를 제공하는 방법을 제안하였다. 제안된 방법은 특허기술문서에서 자주 등장하는 핵심키워드가 특허문서의 기술적 내용을 담고 있는 특징이 있다고 가정하고 있으며, 문서간 핵심어 비교를 통한 유사문서를 검색하였다. 정확도(Precision)측정을 위하여 서로 다른 기술군이 속해 있는 집합에서 질의문서와 유사한 기술군을 분류해보는 실험과 동일한 기술 군에서 질의문서와 유사한 세부기술문서를 찾아보는 실험을 하였다.

본 연구에서처럼 질의문서 자체를 Query(q) 로 사용하여 검색하는 방법은 검색자로 하여금 특허문서 검색에 편의성을 제공한다. 또한 질의문서와 유사한 문서를 검색결과 리스트를 상위에 나타나게 함으로써 특허정보검색의 효율성을 높이는 장점이 있다. 즉, 질의문서와 얼마나 유사한지 정도를 랭킹으로 제공할 수 있게 됨에 따라 유사정도에 대한 정량적인 정보도 줄 수 있으며, 검색을 하는 단계에서 유사한 문서를 우선 검색할 수 있게 하는 기회 등 특허정보검색에 대한 다양한 활용이 이루어 질 것으로 기대된다.

본 연구 한계점은 특허기술문서의 핵심키워드 추출과정에서 의미를 가지지 않는 단어인 관사, 조사, 접속사, 부사, 전치사 등이 포함되어 있어 정확도를 떨어드리는 한계를 지니고 있다. 문서와 문서간의 관계를 분석하여 문서의미에 영향을 미치지 않는 불필요한 단어제거방법이 정교하게 이루어져야 할 것이다. 따라서 문서의미를 대표할 수 있는 단어 추출 연구를 추가적으로 수행할 예정이다.

감사의 글

◆ 이 논문은 2011년도 두뇌한국 21사업에 의하여 지원되었음.
◆ 이 논문은 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임.
(한국연구재단-R1A4A007-2010-0024163)

참고문헌

[1] J. H. Lee, S. H. Cheon, “Re-ranking for Search result using association relationship and TF*IDF”, 한국컴퓨터종합학술대회 논문집, Vol. 37, No. 1, pp. 349-352, 2010.
[2] J. B. Baik, S. M. Kim and S. W. Lee, “Extracting Alternative Word Candidates for Patent Information Search”, 정보과학회 논문지, Vol. 15, No. 4, pp. 299-303, 2009.
[3] S. J. Park and K. T. Kim, “Automatic classification of Web Documents Using Concept-Based Keyword Information”, 한국정보과학회 논문집, Vol. 30, No. 2, pp. 151-153, 2005.
[4] J. M. Koo, S. S. Park, Y. G. Shin. and D. S Jang, “Prediction Method using Text Mining: Focus on Possibility of Patent Registration”, Fall Conference of Korean Institute of Industrial Engineering, 2008.
[5] J. H. Kang, “Information retrieval in a vector space model”, 한남대학교 대학원, 2006.
[6] McGraw-Hill, Salton G. and McGill, M. J. “Introduction to modern information retrieval”, 1983.
[7] WIKIPEDIA, http://en.wikipedia.org/wiki/ Vector_space_ model